Causal Inference
Theory and Applications in Enterprise Computing

Christopher Hagedorn, Johannes Huegle, Dr. Michael Perscheid

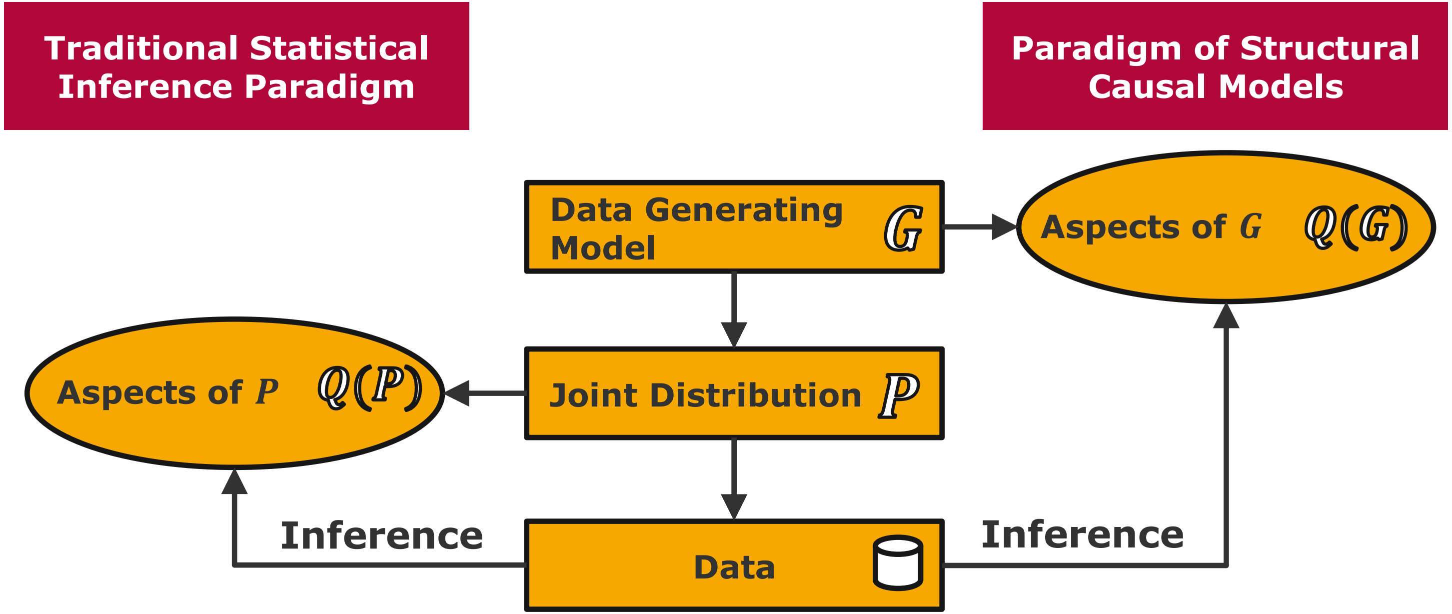
May 12, 2020

Agenda
May 12, 2020

- **Embedding: Causal Inference in a Nutshell**
- **Introduction to Conditional Independence Tests**

Embedding: Causal Inference in a Nutshell

Embedding: Causal Inference in a Nutshell
Concept



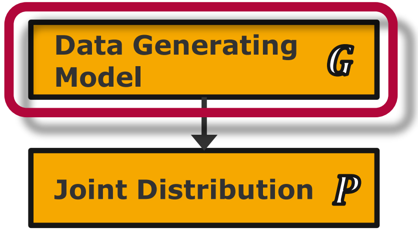
E.g., what is the sailors' probability of recovery when **we see** a treatment with lemons?

$$Q(P) = P(recovery|lemons)$$

E.g., what is the sailors' probability of recovery if **we do** treat them with lemons?

$$Q(G) = P(recovery|do(lemons))$$

Recap: Causal Inference in a Nutshell
Causal Graphical Models



Causal Graphical Model

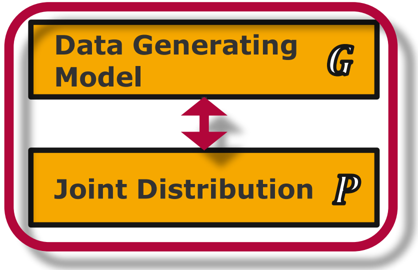
- *Directed Acyclic Graph (DAG)* $G = (V, E)$
  - *Vertices* $V_1, \dots, V_n$
  - *Directed edges* $E = (V_i, V_j)$, i.e., $V_i \rightarrow V_j$
  - *No cycles*
- *Directed Edges* encode direct causes via
  - $V_j = f_j(\mathrm{Pa}(\mathrm{V_j}), \mathrm{N_j})$ with independent noise $N_1, \dots, N_n$

Causal Sufficiency

- All relevant variables are included in the DAG G

Recap: Causal Inference in a Nutshell

Connecting G and P



$$(X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P$$

- Key Postulate: *(Local) Markov Condition*
- Essential mathematical concept: *d-Separation*
  - Idea: *Blocking* of paths
  - Implication: *Global Markov Condition*

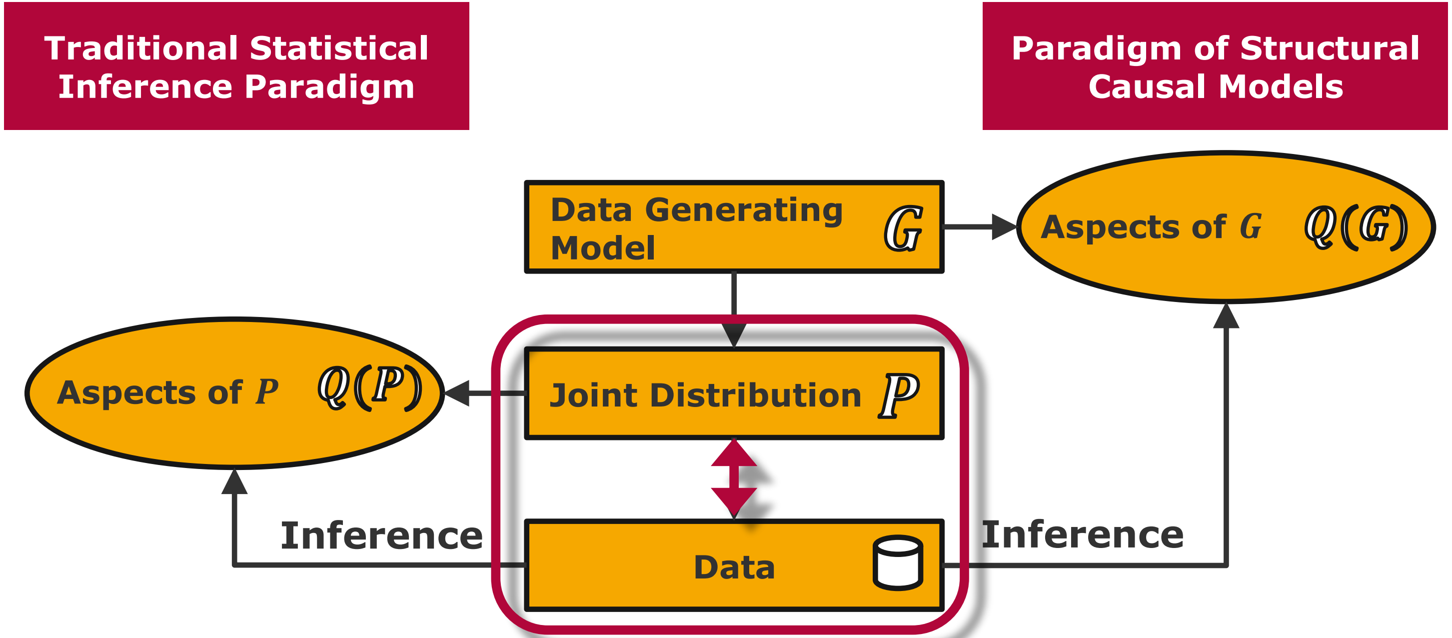
$$(X \perp\!\!\!\perp Y|Z)_G \Leftarrow (X \perp\!\!\!\perp Y|Z)_P$$

- Key Postulate: *Causal Faithfulness*

Introduction to Conditional Independence Tests

Introduction to Conditional Independence Tests
Content

1. Preliminaries
Statistical Inference: Draw Conclusion on P from Data



E.g., what is the sailors' probability of recovery when **we see** a treatment with lemons?

$$Q(P) = P(recovery|lemons)$$

E.g., what is the sailors' probability of recovery if **we do** treat them with lemons?

$$Q(G) = P(recovery|do(lemons))$$

Statistical Inference:
Deduce properties of a population's probability distribution P on the basis of random sampling.

Random samples $X_1, \dots, X_n$

- *independent and identically distributed (i.i.d.)* random variables $X_1, \dots, X_n$

Statistic T

- Function $g(X_1, \dots, X_n)$ of the observations in a random sample $X_1, \dots, X_n$
- Is a random variable with probability distribution (*sampling distribution*)

Point estimator $\widehat{\Theta}$

- Statistic to estimate a *population parameter* Θ

Examples:
Sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ with value $\bar{x}_n$ is an estimator of the population mean μ

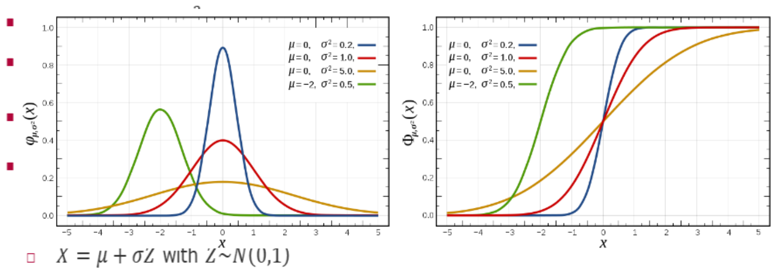
1. Preliminaries
Normal Distribution

Normal Distribution:

We say a random variable X has a normal distribution with mean μ and standard deviation σ if its density function f is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad x \in \mathbb{R}.$$

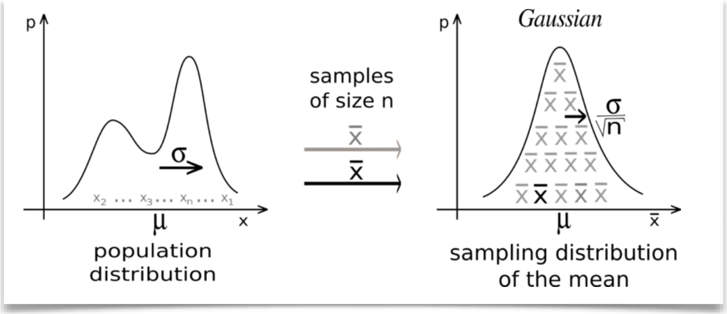
- We write $X \sim N(\mu, \sigma^2)$
- $\Phi_{\mu\sigma^2}(x) = F_X(x) = Pr(X \leq \mathrm{x})$ is the *cumulative distribution function*
- $X \sim N(0,1)$ with $\mathrm{f(x)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ is called *standard normal distributed*
- If $X \sim N(\mu, \sigma^2)$, then
  - $\frac{X-\mu}{\sigma} \sim \mathrm{N(0,1)}$ (*Standardization*)
  - $X = \mu + \sigma Z$ with $Z \sim N(0,1)$



Central Limit Theorem:

For a random sample $X_1, \dots, X_n$ of size n from a population with mean μ and finite variance σ^2 then, for $n \to \infty$,

$$Z = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \to N(0,1).$$



- Therefore, $\bar{X}_n$ is approximately normal distributed with mean μ and standard deviation $\sigma/\sqrt{n}$, i.e., $\bar{X}_n \sim N(\mu, \sigma^2/n)$
- Hence, for the sum $S_n = \sum_{i=1}^{n} X_i$ we have $\mathrm{S_n} \sim N(n\mu, n\sigma^2)$

1. Preliminaries
Confidence Intervals (I/II)

Confidence Interval:
A confidence interval estimate for the mean μ is an interval of the form

$$l \leq \mu \leq u,$$

With endpoints l and u computed from $X_1, \dots, X_n$.

- Suppose that $Pr(l \leq \mu \leq u) = 1 - \alpha,\ \alpha \in (0,1)$. Then for $l \leq \mu \leq u$:
  - l and u are called *lower-* and *upper-confidence bounds*
  - $1 - \alpha$ is called the *confidence level*
- Recall that $\overline{X}_n \sim N(\mu, \sigma^2/n)$. For some positive scalar value $z_{1-\alpha/2}$ we have
  - $\Pr\left(\overline{X}_n \leq \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \Pr\left(\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}\right) = \Phi_{0,1}(z_{1-\alpha/2})$
  - $\Pr\left(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X}_n\right) = 1 - \Phi_{0,1}(z_{1-\alpha/2})$

1. Preliminaries
Confidence Intervals (II/II)

- Therefore

$$\Pr\left(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X}_n \le \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 2\,\Phi_{0,1}(-z_{1-\alpha/2})$$

- Recall, we want

$$\Pr\left(\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X}_n \le \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- With $\alpha = 2\Phi_{0,1}(z_{1-\alpha/2})$ the $100(1-\alpha)\%$ confidence interval on μ is given by

$$\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Since $\alpha = 2\Phi_{0,1}(-z_{1-\alpha/2})$, we can choose $z_{1-\alpha/2}$ as follows:
  - $99\% \Rightarrow \alpha = 0.01 \Rightarrow \Phi_{0,1}(-z_{1-\alpha/2}) = 0.005 \Rightarrow z_{1-\alpha/2} = 2.57$
  - $95\% \Rightarrow \alpha = 0.05 \Rightarrow \Phi_{0,1}(-z_{1-\alpha/2}) = 0.025 \Rightarrow z_{1-\alpha/2} = 2.32$

2. Statistical Hypothesis Testing
Introduction

Knowing the sampling distribution is the key of statistical inference:

- *Confidence intervals*

  Framework to derive error bounds on point estimates of the population distribution based on the sampling distribution
- *Hypothesis testing*

  Methodology for making conclusions about estimates of the population distribution based on the sampling distribution

Statistical Hypothesis:

Statement about parameters of one or more populations

- *Null Hypothesis* H_0 is the claim that is initially assumed to be true
- *Alternative Hypothesis* H_1 is a claim that contradicts the H_0

- A *hypothesis test* is a decision rule that is a function of the test statistic.
  E.g., reject H_0 if the test statistic is below a threshold, otherwise don't.

2. Statistical Hypothesis Testing
Hypothesis Types and Errors

For some arbitrary value μ_0

- *one-sided hypothesis test:*
  $$H_0: \mu \geq \mu_0 \;\; vs \;\; H_1: \mu < \mu_0$$
  $$H_0: \mu \leq \mu_0 \;\; vs \;\; H_1: \mu > \mu_0$$
- *two-sided hypothesis test:*
  $$H_0: \mu = \mu_0 \;\; vs \;\; H_1: \mu \neq \mu_0$$

	H_0 **is true**	H_0 **is false (**H_1 **is true)**
Retain H_0	OK	Type II error
Reject H_0	Type I error	OK

- *Significance level of the statistical test*
  $\alpha = \Pr(\text{type I error}) = Pr(\text{reject } H_0 |\, H_0 \text{ is true})$
- *Power of the statistical test*
  $\beta = \Pr(\text{type II error}) = Pr(\text{retain } H_0 |\, H_1 \text{ is true})$
- *Hypothesis testing*
  Desire: α is low and the power $(1-\beta)$ as high as can be

2. Statistical Hypothesis Testing
Critical Value

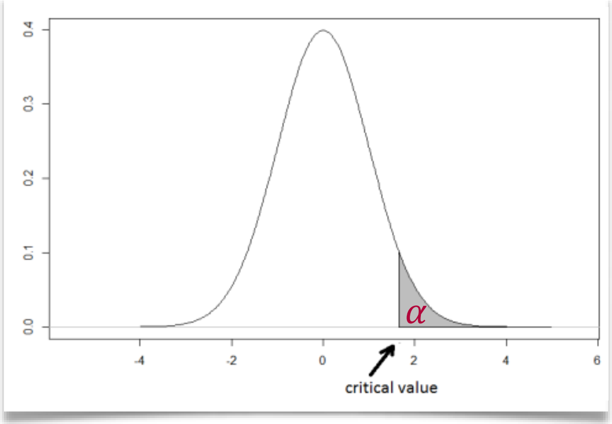
- Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ (σ is known)
- We would like to test $H_0: \mu = \mu_0 \;\; vs \;\; H_1: \mu > \mu_0$

Goal:
Decision rule, i.e., reject $H_0: \mu = \mu_0$ if $\bar{x}_n >$ c for a $c \in \mathbb{R}$

- Choose test statistic T to be $\overline{X}_n$
- Under H_0, we have $T \sim N(\mu_0, \sigma^2/n)$
- $\alpha = P_{\mu_0}(\overline{X}_n > c) = P_{\mu_0}\left(\frac{\sqrt{n}(\overline{X}_n - \mu_0)}{\sigma} > \frac{\sqrt{n}(c-\mu_0)}{\sigma}\right)$

$$= P_{\mu_0}\left(Z > \frac{\sqrt{n}(c-\mu_0)}{\sigma}\right) = 1 - \Phi_{0,1}\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right)$$

- Therefore, $c = \mu_0 + \Phi_{0,1}^{-1}(1-\alpha)\frac{\sigma}{\sqrt{n}}$

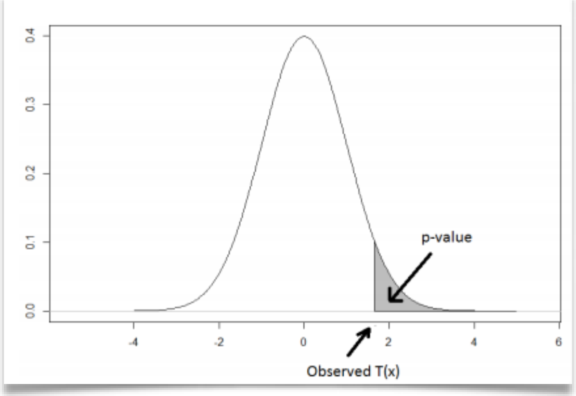


2. Statistical Hypothesis Testing
P-Value

The *p-value* is the probability that under the null hypothesis, the random test statistic takes a value as extreme as or more extreme than the one observed.

- Rule of thumb: p-value low $\Rightarrow H_0$ must go
- We would like to test $H_0\colon \mu = \mu_0 \;\; vs \;\; H_1\colon \mu > \mu_0$
- Here, the p-value is $P_{H_0}(\overline{X}_n > \overline{x}_n) = \cdots$

$$= P_{H_0}\left(Z > \frac{(\overline{X}_n - \mu_0)}{\sigma/\sqrt{n}}\right) = 1 - \Phi_{0,1}\left(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}\right)$$



➡ If $P_{H_0}(\overline{X}_n > \overline{x}_n) < \alpha$ we reject $H_0\colon \mu = \mu_0$

- Absolutely identical to the usage of the critical value

2. Statistical Hypothesis Testing
Supplement: Z-Test

- If the distribution of the test statistic T under H_0 can be approximated by a normal distribution the corresponding statistical test is called *Z-test*
- Overview for Z-tests with known σ:

Testing Hypotheses on the Mean, Variance Known (Z-Tests)

Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with μ unknown but σ^2 known.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}, \quad Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$.

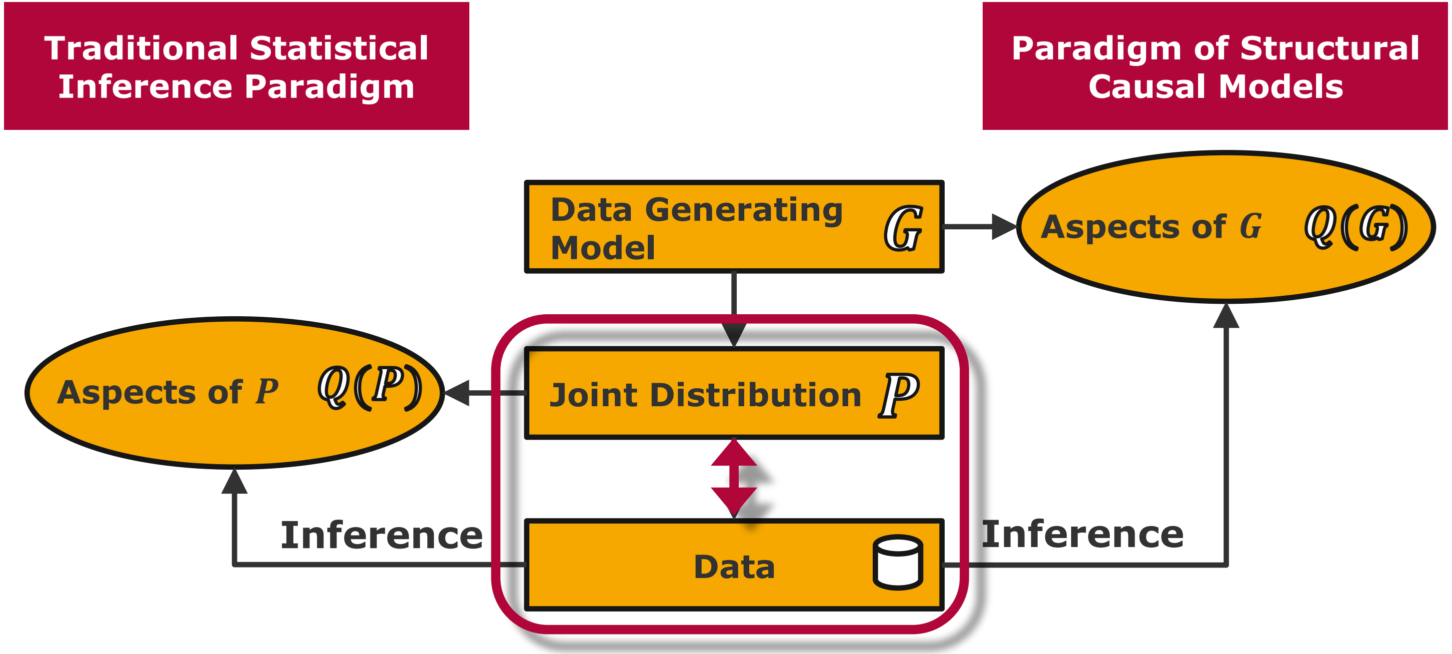
Alternative Hypotheses	P-value	Rejection Criterion for Fixed-Level Tests
$H_1 : \mu \neq \mu_0$	$P = 2[1 - \Phi(\lvert z \rvert)]$	$z > z_{1-\alpha/2}$ or $z < z_{\alpha/2}$
$H_1 : \mu > \mu_0$	$P = 1 - \Phi(z)$	$z > z_{1-\alpha}$
$H_1 : \mu < \mu_0$	$P = \Phi(z)$	$z < z_{\alpha}$

2. Statistical Hypothesis Testing
Summary

- Hypothesis
  - *Null Hypothesis* H_0 is the claim that is initially assumed to be true
  - *Alternative Hypothesis* H_1 is a claim that contradicts H_0
- *Hypothesis test* is a decision rule that is a function of the test statistic T
- How to test a hypothesis?
  - Relation test and confidence interval
  - Approximate T under H_0 by a known sampling distribution $P_{H_0}(T)$
  - Different distributions yield to different tests, e.g., T-test, χ^2-test, etc.
  - Derive rejection criteria for H_0
    - c-value: reject H_0 if $T(x_n) > \text{c}$ for a $c \in \mathbb{R}$
    - p-value: reject H_0 if $P_{H_0}\big(T(X) > T(x)\big) < \alpha$

    are equivalent

3. (Conditional) Independence Testing
Concept (I/II)



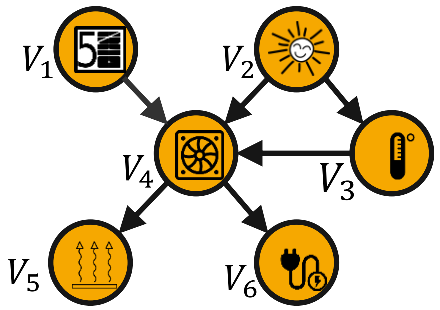
➡ **Use statistical hypothesis tests to obtain information about $(X \perp\!\!\!\perp Y \mid Z)_P$**

3. (Conditional) Independence Testing
Concept (II/II)

Basic idea:
Find a measure T of (conditional) dependence within the random samples $X_1, \dots, X_N$ and apply statistical hypothesis tests whether $T(X_1, \dots, X_N)$ is zero or not, i.e.,

$$H_0: t = 0 \;\; vs \;\; H_1: t \neq 0$$

Cooling House Example:



$V_1, \dots, V_N$ multivariate normal

Correlation coefficient

$$\rho_{V_i,V_j} = cor(V_i, V_j) = \frac{cov(V_i, V_j)}{\sigma_{V_i}\sigma_{V_j}}$$

as measure of linear relationship

3. (Conditional) Independence Testing
Multivariate Normal Data (I/II)

Theorem:
Two bi-variate normal distributed variables V_i and V_j are *independent* if and only if the correlation coefficient ρ_{V_i,V_j} is zero.

- Hence, we test whether the correlation coefficient ρ_{V_i,V_j},

$$\rho_{V_i,V_j} = \frac{E\left[\left(V_i - \mu_{V_i}\right)\left(V_j - \mu_{V_j}\right)\right]}{\sigma_{V_i}\sigma_{V_j}},$$

  is equal to zero or not, i.e., $H_0\colon \rho_{V_i,V_j} = 0 \quad vs \quad H_1\colon \rho_{V_i,V_j} \neq 0$

- For i.i.d. normal distributed V_i, V_j, applying Fisher's Z-transformation ρ_{V_i,V_j},

$$Z\left(\rho_{V_i,V_j}\right) = \frac{1}{2}\log\left(\frac{1+\rho_{V_i,V_j}}{1-\rho_{V_i,V_j}}\right),$$

  yields to $Z\left(\rho_{V_i,V_j}\right) \sim \mathrm{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho_{V_i,V_j}}{1-\rho_{V_i,V_j}}\right), \frac{1}{\sqrt{n-3}}\right)$.

3. (Conditional) Independence Testing
Multivariate Normal Data (II/II)

- Thus, we can apply standard statistical hypothesis tests, i.e.,
  - Derive p-value

$$p(V_i, V_j) = 2\left(1 - \Phi_{0,1}\left(\sqrt{n-3}\,\left|Z\left(\rho_{V_i,V_j}\right)\right|\right)\right)$$

  - Given significance level α, we reject the null-hypothesis $H_0\colon \rho_{V_i,V_j} = 0$ against $H_0\colon \rho_{V_i,V_j} \neq 0$ if for the corresponding estimated p-value it holds that $\hat{p}(V_i, V_j) \leq \alpha$
- This can be easily extended for conditional independence:

Theorem:
For multivariate normal distributed variables $\boldsymbol{V} = \{V_1, \dots, V_N\}$ we have that two variables V_i and V_j are conditionally independent given the separation set $\boldsymbol{S} \subset \boldsymbol{V}/\{V_i, V_j\}$ if and only if the partial correlation $\rho(V_i, V_j | \boldsymbol{S})$ between V_i and V_j given $\boldsymbol{S}$ is equal to zero.

- I.e., we can apply the same procedure to receive information about conditional independencies

- Statistical hypothesis testing theory allows to obtain $(X \perp\!\!\!\perp Y | Z)_P$ from data
- Distribution of $V_1, \ldots, V_N \Rightarrow$ dependence measures $T(V_i, V_j, \boldsymbol{S}) \Rightarrow$ hypothesis test $H_0: t = 0$

Examples

- Multivariate normal data:

$$Z(v_i, v_j | \boldsymbol{s}) = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_{v_i, v_j | s}}{1 + \hat{\rho}_{v_i, v_j | s}} \right)$$

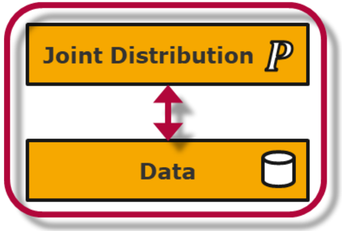
with sample (partial) correlation coefficient $\hat{\rho}_{v_i, v_j | s}$

- Categorical data:

$$\chi^2(v_i, v_j | \boldsymbol{s}) = \sum_{v_i v_j s} \frac{\left(N_{v_i v_j s} - E_{v_i v_j s}\right)^2}{E_{v_i v_j s}} \quad \text{and} \quad G^2(V_i, V_j | \mathrm{S}) = 2 \sum_{v_i v_j \boldsymbol{s}} N_{v_i v_j s} \ln \left(\frac{N_{v_i v_j s}}{E_{v_i v_j s}} \right)$$

with $E_{v_i v_j s} = \frac{N_{v_i + s} N_{+ v_j s}}{N_{++s}}$ *where* $N_{v_i +} = \sum_{v_j} N_{v_i v_j}, N_{v_i +} = \sum_{v_j} N_{v_i v_j},$
$N_{+v_j} = \sum_{v_i} N_{v_i v_j}$ *and* $N_{++} = \sum_{v_i v_j} N_{v_i v_j}$ *are calculated for every realization of* $\boldsymbol{S}$

- This defines the basis of constraint-based causal structure learning



References

Literature

- Lehmann et. al. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.
- Montgomery et. al. (2010). *Applied statistics and probability for engineers*. John Wiley & Sons.
- Dempster et. al. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley Publ. Co., Reading, Mass. 1969.
- Joe Whittaker. (2009). *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing.

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Thank you
for your attention!